I J C T A, 10(8), 2017, pp. 369-376 ISSN: 0974-5572

Large Graph Analysis and Visualization in Graph Mining: A Survey

Swati K. Bhavsar* and Varsha H. Patil**

ABSTRACT

Large graph is used for storing the information and it is a very complex data structure. Hence there is a need of analyzing and visualizing such graphs, which is performed in graph mining. The Graph mining finds applications are in the field of bioinformatics, chemical reactions, Program flow structures, computer networks, social networks etc. Also it is useful in data mining and mined data analysis.

In literature of this paper various graph mining approaches along with graph analysis have been discussed. These are based on different classification techniques, decision trees and clustering methods. This paper also discusses graph databases which are available to analyze the performance of graph analysis techniques.

Keywords: Large Graph, Analysis, Visualization

1. INTRODUCTION

Large graph is made up of hundreds to thousands of nodes and millions of edges. Examples of large graphs are Web graphs, social networks, recommendation system. To find patterns in large graph, it is desirable to analyze, visualize, summarize and mine it. As it is a complex data structure such graphs require excessive processing, more memory for storage and knowledge of a pattern of the graph. Some of the graphs changes with time are known as dynamic graphs. And hence it is very difficult to comment on exact size and pattern of such large graphs. It leads to analysis of graphs. Those are discussed under graph mining.

Graph mining techniques are categorized into various groups as follows.

(1) Graph clustering: The vertices of the graph are grouped into various clusters by considering edge structure such that many edges within the cluster and few edges between the cluster. And the graph clusters are formed based on similarity of structure of graph. It is based on unsupervised learning technique as classed are not known prior to the clustering.

(2) Graph Classification: It deals with the separation of graph database into two or more classes. It may be based on supervised or unsupervised learning techniques.

(3) Sub graph mining: For a large graph, many sub graphs can be created. And the vertices and edges are subsets of another graph. To divide the large graph into sub graphs, it needs partition of graph is known as sub graph mining [1]. To find out sub graph and to achieve minimum loss of information while partitioning, its important to find threshold These are focused as graph analysis, Graph Visualization and Graph Summarization.

Large Graph Analysis: It deals with partitioning graph into sub graphs

Large Graph Visualization: It deals with finding data elements and their internal relationship into graph.

Large Graph Summarization: It deals with finding the stronger connected component.

* Research Scholar : Dept. of Computer Engg., MCERC, *Email: Nashikbhavsar.swati4@gmail.com*

** Vice Principal., MCERC Nashik, *Email: varsha.patil@gmail.com*

2. Basic Graph Theory

A graph G is a set G = (V, E). where, V is the set of vertices. E contains the set of edges of the graph and the number of vertices n = IVI is the order of the graph. In an undirected graph, each edge is an unordered pair {v, w}. There are several terms related to the graph as discussed below.

- Planer Graph: It is a graph drawn in a plane without any of the edges crossing.
- Connected Graph: It is a graph drawn, If there exist paths between all pairs of vertices. The graph is called as disconnected, if the vertices cannot be reached from each others. The interconnection of graph is maintained through edge connectivity. Graph connected Graph G can be converted into disconnected Graph by removing minimum number of edges is also referred as edge connectivity of graph.
- Cyclic Graph: If a graph contains a cycle means a simple path that begins and ends with the same vertex. And if a graph that contains no cycle then it becomes acyclic and is also known as a forest. A connected forest is termed a tree.
- Graph Sparsity / Density: Graph Sparsity of a graph is a graph in which the number of edges is close to the minimal number of edges. Graph density of a graph in which the number of edges is close to the maximal number of edges.
- Static / Dynamic : Static graph consist of a fixed number of nodes and edges. While, in a Dynamic graph, insertion and deletion of edges and vertices can be performed at any time
- Weighted / Unweighted Graph Distribution: It is a graph in which each branch is given a numerical weight. Therefore it is a special type of labelled graph in which labels are numbers, where unweighted graph is a graph in which weights are not assigned to branches.
- Vertex Degree Distribution: Degree of a network is the number of connections and vertex degree distribution is probability distribution of these degrees.
- Directed / Undirected: It is a set of nodes or vertices that are connected together, where all the edges are directed from one vertex to another. It is sometimes called as digraph. Whereas vertices are not directed from one another is known as undirected graph.
- Simple/ Multi graph: simple graph is an un weighted, undirected graph containing no graph loops or multiple edges. Where multi graph is an undirected graph in which multiple edges are allowed. It may have multiple loops as well.

3. LITERATURE REVIEW

This section summarizes the different proposed graph mining and analysis methods.

Callut et al. have proposed a new technique called D-Walks. Dwalks classifies the unlabelled nodes are compared by using betweenness parameter. The CORA database is used for its implementation. Different experiments have been performed on it. And the experimentation proves that it gives good results for semi supervised classification of large graphs. [2].

Kasima et.al presented a method based for large graph classification based on kernel method [3]. The inner product of two graph is computed, further feature space is created for the classification of graphs. Unknown graphs are taken as an input and classified into particular class. The similarity of two graphs is computed depending upon nodes of the graphs and labels of edges of the graphs. If the similarities of graphs are identical, then they are classified into same group. It is useful for the prediction of chemical compound characteristics. The datasets used are mutag and PTC.

Dhillon et al. have presented an efficient and fast technique for graph clustering. By using multilevel approach clusters of nodes are formed. on the IMDB Movie dataset is used for implementation. The dataset has 1.2 million nodes and 7.6 million edges. Furthermore different experiments have been performed using this dataset. These techniques compute 5000 cluster and coarsening of the nodes performed by using multi eigenvectors [4].

Le et al. have proposed method for clustering of bi-partite graph. It is also called as Coring technique. It is mainly based on partitioning a large graph into small sub graphs. The nodes of the clustered sub graphs are strongly interconnected within graph and weakly connected to the nodes of other graphs. Their method is called coring method, which is able to handle both weighted and unweighted graph. It works in following steps[5]

Step 1: In this step, the coring method is used and it computes the density variation sequence of nodes . The method iteratively computes the minimum density D and set of nodes having minimum density M. The output of this step is sequence of D,s and M,s.

Step 2: It this step, the core nodes in the given graph are identified. The rate of decrease/ increase in value of minimum density is computed. If the rate of increase/decrease in the D value is greater than the threshold and the sequence of M is also in some order then the nodes are identified as core nodes.

Step 3: In this step, the large graph nodes are partitioned into clusters.

Step 4: In this step, core groups are expended into different clusters.

By using this technique, the high density nodes are placed at center and lower density nodes are encircling these nodes. It has been implemented on microarray dataset. This dataset contains some of tumor as well as normal tissues. It is able to clusters the tumor tissues as well as normal tissues.

Karypis and Kumar has proposed k-way graph partitioning known as METIS [8] which is based on multilevel partitioning. In this method, the graph size is reduced by applying graph partitioning. The vertices and edges are broken up, it is also known as coarsening method. The large graph is partitioned into smaller sub graph. And uncoarsen is applied to construct a partition for the original graph. As the partitions are stored in adjacency matrix, which is a f iixed data structure, Run time addition or deletion in a sub graph is not possible. This is major limitation of this approach.

Chao – Wei Ou and Sanjay Ranka has suggested Parallel Incremental Graph Partitioning using Linear Programming [9] which is used to execute several scientific and engineering applications parallel, requires the partitioning of data or among processors to balance computational load on each node with minimum communication. There are many algorithms like geometric, structural, spectral & refinement algorithms are proposed for achieving parallel graph partitioning.

Stephen Barnad, Horst D. Simon has suggested Fast multilevel implementation of recursive spectral bisection for partitioning unstructured problems which is used for the partitioning of the graph which needs to be updated as the graph changes over time i.e a small number of nodes or edges may be added or deleted at any given instant. The drawback of the method is initial partition is to be calculated using linear programming based bisection method. [10] The Proposed approach focuses on uniform partitions creation with no loss of information.

Inderjit S. Dhillon, Yuqiang Guan has described an [11] equivalence between the objective functions and high quality multilevel algorithm that optimizes various weighted graph clustering objectives such as ratio cut, normalized cut and ratio association criteria.

Mahmudur Rahman, Mansurlal Alam Bhuiyan discusses and efficient graphlet counting method for large graph analysis. Which computes the cost for obtaining frequency of each graphlet in the network. [12] Also the local topological structure is considered for the computation of the Graph Frequency Distribution (GFD).

Vladimir Batagelj, Franz J. Bran denburg, Water Didimo described (X, Y)-clustering and Hybrid visualizations technique for visual analysis of large graphs In (X,Y) clustering the two important properties like intra cluster and inter cluster were used for topological graph. [13] By using hybrid visualizations clusters were able to explore the clusters without losing their mental map.

Jose F. Rodrigues, et. Al had discussed Large Graph Analysis in the GMine System for the clutter reduction in the graph which is based on graph representation as hierarchies of partition by using concept of super graph and sub graph. And graph summarization is performed using Center Piece summarization. Their main contribution is for the large graph investigation in terms of locally and globally [14].

4. LARGE GRAPH VISUALIZATION

The graph is seen with thorough graph hierarchy is called as graph visualization. Graph visualization is more powerful with concern to various reasons as graph is a very simple models. For example the World Wide Web, in which nodes are acting as web pages and links can represent hyperlinks among them. Along with it, there are many example, some of them are Computer file system, interpersonal relationships and animal spices tress. But it needs to process the graph efficiently and effectively. For this purpose the graphs are compared with different common visual Design principles such as symmetry, relative size, similarity and closure. So the various layouts of the graphs are needed to discuss.

4.1. Graph Layout

The graph can be represented in one of the following layouts.

- Node-Link Layout: The parent child relationship can be indicated by links between the nodes. For the node link layout representation some of the points are to be considered such as nodes and edges should be evenly distributed, the crossing of edges should be avoided, the minimization of bends of edges and to Maintain the isomorphic substructure of the graph. It is further classified into Tree Layout, Tree+ link layout and spring layout.
- Tree Layout: The nodes of the graph are placed as shown in the figure 1 as parent node is at the root node and child nodes as intermediated nodes of the tree structure. But this representation has a major problem that the unused space is more and root side and becomes denser at the opposite side.

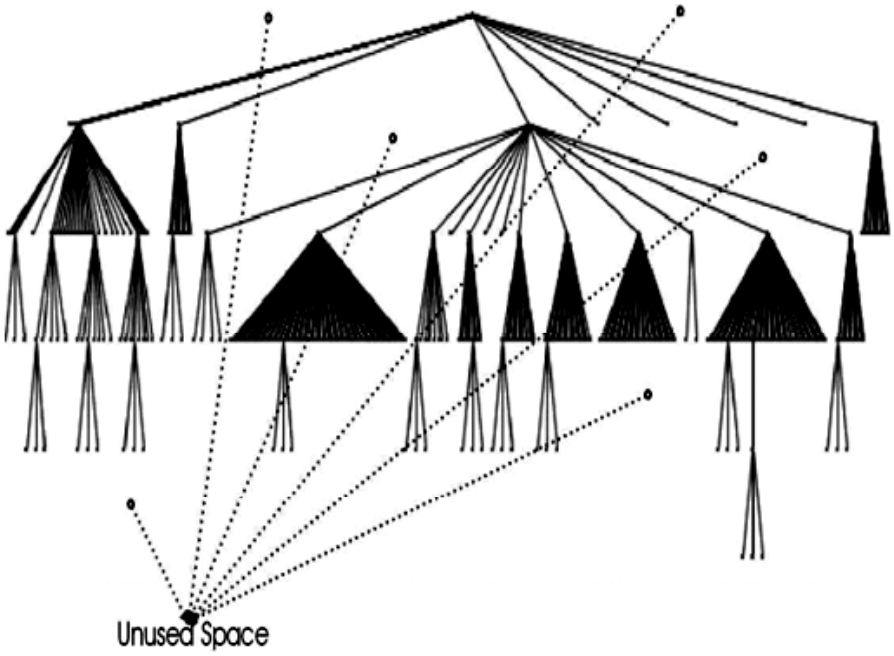

Figure 1: Tree Layout

The solution to this problem can be provided by node link layout as shown in figure 2. It places the children of a sub tree into a circular sedge shape according to their depths in a tree. The root node known as a focus node , which is placed at a center. The other nodes are radiating outward place on a separate circles.

Figure 3 shown the balloon layout in which it follows the cone tree structure where father node is at center it is surrounded y siblings of a sub tree.

Various graph visualization techniques are as follows.

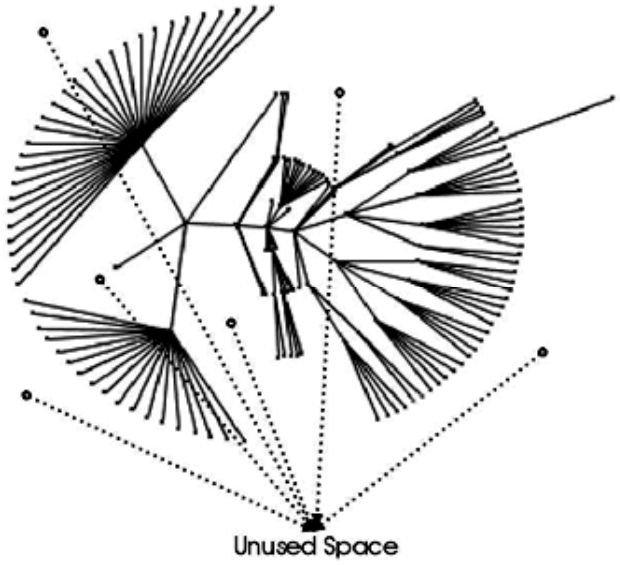

Figure 2: Node Link Layout

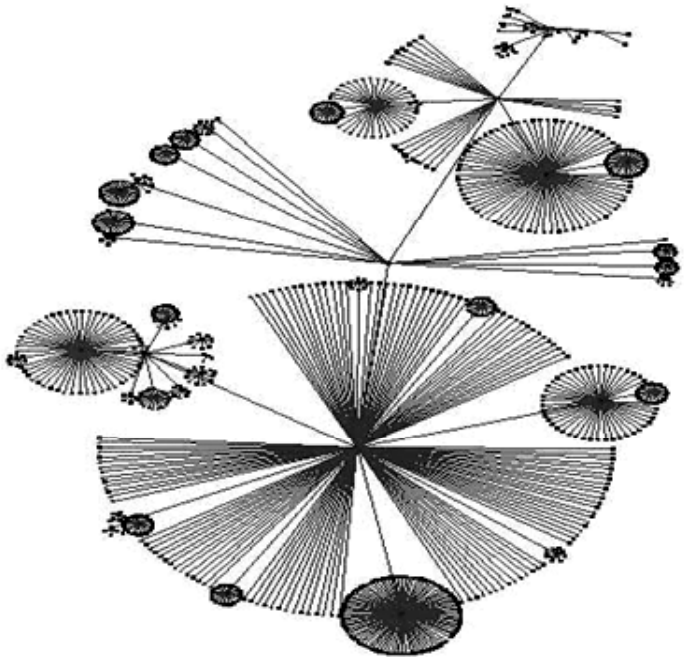

Figure 3: Balloon Layout

4.2. Visual Clutter Reduction

If the large graph contains more no. crossed edges and it cannot be seen through a single glance, then it leads to dividing into sub graphs, then some of the information get lost in terms of edges and vertices. To avoid this widely used ways are edge displacement, node clustering and sampling [15].

- *Edge Displacement*: The visual clutter can be reduced by drawing the edges in different shapes such as splines and polylines and lso by reducing the cross edges. Edge drawing is important in calculating geographical locations where preassigned positions are calculated. But to find optimized solutions to the large graph is a very time consuming process.
- *Node Clustering*: When a large graph is divided onto sub graph structure, then it is called as clustering. Similar visual elements are grouped together so that more free space is released. And it will definitely help in reducing the visual clutter. This is further classified into following types.
- *Graph theoretical*: These algorithms find out the similarity between individual nodes and represented using similarity matrix. Furthermore closely related nodes depending upon the threshold value are grouped together to form a cluster. Each cluster is a connected graph which uses various bisection algorithms such as spectral bisection and multilevel bisection.
- *Single-pass*: First of all cluster seeds as an individual data points are find out. And then the clusters are formed by growing them. For example, a starting node has been chosen and then nodes are added one by one until cluster becomes big. At each step, greedy algorithms are applied for finding which node is to be added.
- *Iterative algorithms*: This algorithm is based on other clustering algorithm. One example of this type of algorithm is hierarchical clustering in which larger cluster is formed by merging small clusters. In hierarchical clustering layers can be produced. And these layers are useful for finding the depth of the graph. Furthermore, by finding the cluster sequence in detail, partition quality of a graph can be improved by transferring information between levels. This method includes the steps lie coarsening, partitioning and projecting.

Some of the properties of good clustering are discussed as blow.

- Balanced cluster: if the size of cluster is same in each level of hierarchy along with nodes distribution is even. Then the cluster is balanced cluster.
- Small cluster depth: If graph contains a small number of layers in the recursive decomposition is known as small cluster depth.
- Convex cluster drawings: The cluster drawing fit in a simple convex region.
- Balanced aspect ratio: Cluster regions should not be too thin.
- Efficiency: Cluster computing should not take too long.
- Symmetry: Display symmetry should be maximum.

4.2.1. Sampling

Sampling is used for the reduction of visual clutter for preserving the topological properties of a network. An important notation “focus” is used and it can be assigned by the user. Therefore visualization depends open the focal area and its neighborhoods in the given large graph. Hence visualization is categorized into the following groups.

- Select: It is useful to highlight certain focus targets.
- Explore: It is useful to change current view point to another part of the data in the same layout representation, such as panning and rotating.
- Reconfigure: It is useful for switching in different layout, such as replacing nodes in graphs
- Encoding: It is used to switch different representation schemes, such as changing the layout from node-link representation to tree maps representation.

- Abstract/Elaborate: It is useful to give users different insight into data by adjusting the levels of data abstraction.
- Filter: It is useful for displaying the data as per user's request
- Connect: It is used to highlight the connections between items.

Along with this the following methods of large graph visualization are very important.

4.2.2. Zoom and Pan

Zooming and panning are fundamental tools for exploring large information. Zooming is useful for detailed insight of data and Panning is smoothly moving camera across scene. They are go together to each other in functionality and quite vital when large graph structures are explored.

- *Filtering*: Filtering refers to hide s items from the view. It is a removal of any data whose attribute values below a threshold. Although the concept of filtering is simple, a useful visual filtering interface should provide various visual browsing tools, such as fast and continuous display of results, progressive refinement of parameters, Ahlberg et al. [18] propose the dynamic query filters for visual information seeking. Their query parameters are adjusted with sliders, buttons and so on. A key to these principles is to understand the enormous capacity for human visual information processing. It also utilizes the powerful perceptual ability of human beings to help rapidly filter the viewing items.

5. GRAPH DATASETS

To analyze the performance of these methods following data sets are used.

1. DBLP Data Set: It is a database of Computer Science publication which represents authors of a graph and their publication details
2. Twitter: as a Social Networks

6. CONCLUSION

In this paper, we have presented information of the different graph mining techniques. These graph mining techniques are based on the classification, clustering, decision tree approaches, which are the data mining fundamentals. In addition we also have presented information for large graph Analysis, Visualization and summarization.

REFERENCES

[1] N. S. Ketkar, L.B. Holder and O. Cook,: Empirical Comparison of Graph Classification Algorithms, IEEE, (2009)

[2] J. Callut, K. Fran 90isse, M. Saerens and P. Dupont, :Semi-supervised Classification from Discriminative Random Walks, Lecture Notes in Artificial Intelligence No. 5211, Springer, pp. 162-177, (2008).

[3] H. Kashima and A. Inokuchi, "Kernels for graph classification", ICDM Workshop on Active Mining (2002) .

[4] Dhillon, Y. Guan and B. Kulis: A Fast Kernel-based MultilevelAlgorithm for Graph Clustering, Proceedings of The 11th ACM SIGKDD, Chicago, IL, Aug. 21-24, (2005)

[5] T. V. Le, C. A. Kulikowaski and I. B. Muchnik :Coring Method for Clustering a Graph, In proceedings of IEEE, (2008).

[6] Y. Chen and F. Fonseca :A Bipartite Graph Co-Clustering Approach to Ontology Mapping , (2004)

[7] G. Karypis, V. Kumar: Multilevel Graph partitioning schemes, Proc. IEEE/ ACM conf. Parallel processing, pp. 113-122, (1995)

[8] Chaw Wei, Ou, Sanjay Ranka: Parallel Incremental Graph Partitioning, IEEE transactions on Parallel and Distributed Systems, Vol. 8, No. 8, (1997).

[9] Stephan T. Barnad, Horst D. Simon, :Fast multilevel Implementation of recursive spectral bisection for partitioning unstructured problems, Concurrency: Practice and Experience,.Vol6 (2) , pp 101-117, (1994).

[10] Inderjit S. Dhilon, Yuqiang Guan, Brian Kulis :Weighted Graph Cuts without Eigenvectors: A Multilevel Approach, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.29, No.11, pp. 1944-1957. November (2007).

[11] Mahmudar Rahman, Mansurul Alam Bhuiyan, Mohammad Al Hassan :GRAFT: An efficient Graphlet counting method for Large Graph Analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.26, No.10, pp. 2466-2478. October (2014).

[12] Vladimir Batagelj, Franz J. Bran denburg, Water Didimo :Visual Analysis of Large Graphs using (X, Y)- clustering and Hybrid Visualizations, IEEE Transactions on Knowledge and Data Engineering, Vol. 17, No. 11, pp. 1587-1598. November (2011).

[13] Jose F. Rodrigues Jr., Jia-Yu Pan, Agma J.M. Traina, Caetano Traina Jr., Christos Faloutsos, :Large Graph Ananlysis in the Gmine System, IEEE Transactions on Knowledge and Data Engineering, pp. 106-119, (2013)

[14] Geoffrey Ellis and Alen Dix. : Taxonomy of Clutter Reduction for Information Visualization, IEEE Transactions on Visualization and Computer Graphics, 13(6):1216, 1223, 2007.

[15] M. R. Garey and D. S. Johnson.: Crossing number is Np-complete. SIAM Journal on Algebraic and Discrete Methods, 4(3):312–316, 1983.